## 总体

目标是100%准确率

尽管可能在不完成5个CheckPoint的情况下达成100%准确率，但是给分的时候可能出问题

## 英文

任务名字叫OCR或HTR，可以作为关键词检索

Checkpoints

1) How do you eliminate the influence of unrelated factors? (Such as lines, noise, etc.)

通过**颜色处理**，尝试使用HIS颜色空间区分，查表等等

检查是否有**noise**，如果有考虑形态学去噪

2) How do you preprocess the text?

预处理，或者使用形态学细化后用连通域分割

**斜过来的角度**可能需要一些hough变换来检测到，考虑先threshold再聚类（再加一些解释，竖向直线在字母中的最显著性）；或者找专业的skew estimators来做

字是斜的，把原图**正过来**然后匹配正模板

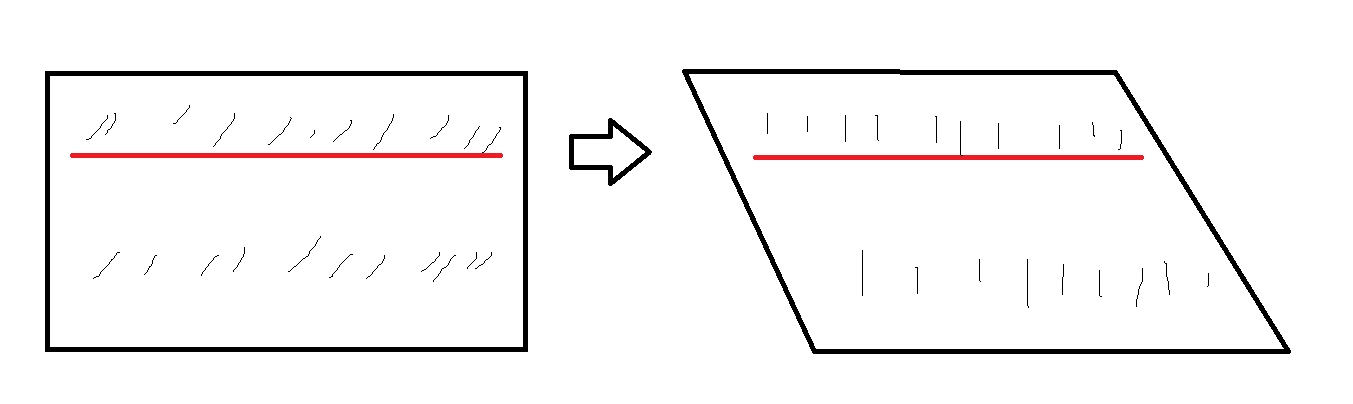
**二值化**

3) How do you segment the characters?

**分行**，每行**分词**，使用histogram（一个维度积起来）；专业的算法是recursive projection，递归投影分割，用于文档版面分割

**分字母**

考虑shift斜过来，然后基于histogram分词的逻辑进行分字母



4) How do you perform text recognition?

法1：**模板匹配**，手写26个英文字母作为模板？，找其它来源的匹配做模板？这个需要快做起来；模板匹配需要找各尺寸模板匹配方法，而不是naïve版本的

法2：找一些斜体的，转正，统计，各种统计量，这个没想好怎么搞，可能需要查一下；或模板匹配

5) How is the final effect? (Show your recognition accuracy.)

**需要码进去ground truth，**并有一些指标的计算：detection查全率，recognition查全率，平均编辑距离等（但是如果识别率100%的话其实没什么问题，可以暂时用字符准确率应付一下）

可能需要额外normalize技巧

## 中文

英文识别的弱化版本，做完英文识别再做（也许

**Checkpoints:**

1. How do you eliminate the influence of unrelated factors? (Such as lines, noise, etc.)

同上，使用颜色区分

1. How do you preprocess the text?

分词，分行，同上，而且是更简单版本

1. How do you segment the characters?

跳过

1. How do you perform text recognition?

收集类似风格，对应的中文字，然后可以直接匹配或者计算histogram

可能试试thinning+交点、孔检测

1. How is the final effect? (Show your recognition accuracy.)

这里的匹配准确率计算可以简单，但仍然需要码进去ground truth

## 计划

5.16完成各自的工作，测试组合起来的性能

分锅

韩嘉祺：

识别

陈醉：

分行分词预处理（校正之后的图片）

安丽华：

颜色